Transcriptome analysis and related databases of Lactococcus lactis

Kuipers, Oscar P.; Jong, Anne de; Baerends, Richard J.S.; Hijum, Sacha A.F.T. van; Zomer, Aldert L.; Karsens, Harma A.; Hengst, Chris D. den; Kramer, Naomi E.; Buist, Girbe; Kok, Jan

Published in:
Antonie Van Leeuwenhoek: International Journal of General and Molecular Microbiology

DOI:
10.1023/A:1020691801251

Publication date:
2002

Antonie van Leeuwenhoek **82:** 113–122, 2002.

Transcriptome analysis and related databases of *Lactococcus lactis*

Oscar P. Kuipers*, Anne de Jong, Richard J.S. Baerends, Sacha A.F.T. van Hijum, Aldert L. Zomer, Harma A. Karsens, Chris D. den Hengst, Naomi E. Kramer, Girbe Buist & Jan Kok
Molecular Genetics, University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute, PO Box 14, 9750 AA Haren, The Netherlands (Author for correspondence; E-mail: o.p.kuipers@biol.rug.nl)

Key words: Lactococcus lactis, genome sequence, transcriptome, DNA-microarrays, gene regulation, functional genomics, bioinformatics

Abstract

Several complete genome sequences of *Lactococcus lactis* and their annotations will become available in the near future, next to the already published genome sequence of *L. lactis* ssp. *lactis* IL1403. This will allow intra-species comparative genomics studies as well as functional genomics studies aimed at a better understanding of physiological processes and regulatory networks operating in lactococci. This paper decribes the initial set-up of a DNA-microarray facility in our group, to enable transcriptome analysis of various Gram-positive bacteria, including a ssp. *lactis* and a ssp. *cremoris* strain of *Lactococcus lactis*. Moreover a global description will be given of the hardware and software requirements for such a set-up, highlighting the crucial integration of relevant bioinformatics tools and methods. This includes the development of MolGenIS, an information system for transcriptome data storage and retrieval, and LactococCyc, a metabolic pathway/genome database of *Lactococcus lactis*.

Introduction

Lactococcus lactis ssp. *lactis* IL1403 was the first lactic acid bacterium for which the complete genome sequence and its annotation have been published (Bolotin et al. 2001). This important achievement has set the stage for various other genome sequencing projects on *L. lactis* and on a wide variety of other lactic acid bacteria (see also Klaenhammer et al. 2002 in this issue). The completion of the genome sequence of *L. lactis* ssp. *cremoris* MG1363 by an English/Irish/Dutch consortium is in progress and is expected to be completed in 2002. Moreover, in the USA an effort is going on to sequence the chromosome and plasmids of *L. lactis* ssp. *cremoris* SK11 (D. Mills et al., personal communication), the sequence of which is also expected to be finished in 2002. It is very likely that the genomes of even more strains will be sequenced in the near future, enabling comparative genomics within this species that is of such great interest to the dairy industry. Next to *Bacillus subtilis*, which is the paradigm for genetic, physiological and biochemical research in Gram-positive bacteria, *L. lactis* has evolved to be a model species for Gram-positive cocci.

The availability of genomic information on lactococcal and other bacterial species has opened the way for a number of analytical and experimental approaches, which were impossible to perform without this data. First, genome mining and comparison studies yield valuable information on the presence or absence of certain features in the strains studied, as well as on evolutionary phenomena. For instance, the IL1403 sequence has revealed the presence of some unexpected pathways, e.g., on competence development and some expected but interesting traits, such as respiration ability (Duwat et al. 2001), as well as on horizontal gene transfer (Bolotin et al. 2001). Apart from the mining approaches, which in essence will primarily place genes in functional categories and allow to compare gene organizations in various related and unrelated species, it is exciting to see how fast a number of experimental approaches have evolved to enable transcriptome analysis, genomotyping, proteome analysis, high-throughput screening and high-throughput structural biology in microbial systems (Kuipers 2000; Paton et al. 2000; Sebaihia et al. 2001; Tomita 2001; Ye et al. 2001; Zheng et al. 2001; Oliver et al. 2002). Since there is a huge interest to identify

lactococci with desired properties or to construct such strains by genetic modification, it is essential to know how the function of each individual gene/protein is related to its expression/production during growth and in stationary phase in various environments. For this reason the elucidation of gene regulatory networks and identification of protein–protein interactions involved in important industrial processes is of crucial importance. Fortunately, these issues can now be addressed in a systematic way, by using the genomics tools that have evolved over the past years. Research using these tools will undoubtedly be directed towards studying important industrial phenomena like acidification rate, various stress responses, flavour formation, effects of metabolic engineering, assessment of the safety of GMOs, use of lactococci as delivery vehicles in a host or food-product, or for improved biopreservation.

One of the most powerful approaches is to use proteomics to determine the full complement of proteins in a cell under a specific condition and at a defined time. Technology development in this field is taking place rapidly, e.g., directed to the development of protein- or antibody arrays and improved and high-throughput protein separation techniques. However, the most commonly used methodology is to combine separation of proteins by 2D-electrophoresis, followed by trypsinolysis and protein identification by MALDI-TOF. Some drawbacks of this approach are the difficulty to detect low-abundancy proteins, and the problems in analysing membrane proteins and unstable proteins.

Another very powerful tool in genomic research is provided by DNA-microarrays, which offer a near-complete view of the relative abundance of all messenger RNAs in a cell in a spatio-temporal way. Various technologies for the production and analysis of DNA-microarrays have been described (Diehl et al. 2001; Kamb & Ramaswami 2001; Lucchini et al. 2001; Wei et al. 2001; Ball & Trevors 2002), but it is beyond the scope of this paper to further describe them. The purpose of this paper is to describe in some detail how DNA-microarray analysis has been set-up in our group, including a description of the infrastructure and the bioinformatics tools that are extremely important for being able to extract useful information from primary transcriptome data. The set-up that will be described has already been implemented for microarrays of *Bacillus subtilis* 168 and for *Streptococcus pneumoniae* TIGR4 (collaboration with Dr. P.W.M. Hermans, Pediatrics, Rotterdam), while arrays for a *Lactobacillus* species and another *Bacillus* species will be developed in the near future in various collaborations with groups having in-depth expertise with these organisms.

DNA-microarray production

Because standardisation of DNA-microarray production and analysis is of crucial importance to be able to create robust and validated data, a series of steps in the whole procedure was carefully analysed and optimised. Amplicons of all genes of the genome studied are synthesised by PCR, have a maximal length of 800 bp and are commonly, if gene size allows, longer than 500 bp. The 5′ and 3′ specific primers are synthesized with generic tags (15 nucleotides forward and reverse, respectively), allowing reamplification of the amplicons obtained in the first round of PCR (original amplicons). These amplicons can be reamplified with an aminated forward primer and a non-aminated reverse primer-based on the generic tag sequences, which yield single-aminated amplicons after PCR. This approach has the advantage of only detecting cDNAs originating from gene transcripts and not those originating from antisense RNA, produced by read-through from genes/operons into downstream genes in a back-to-back orientation.

The purity, size-correctness and amounts of the amplicons are routinely checked by running them on agarose gels. Subsequently, amplicons are purified by the Millipore 96 colums PCR purification system. Amplicons are transferred to 384-well microtiter plates and are diluted 1:1 with Array-It spotting solution to a final concentration of 0.25–1.0 mg/ml. Using Telechem stealth pins (SMP5) 0.2 μl is sampled and 1–3 nL are spotted in duplicate onto aldehyde-coated slides (Cell Associates) and further handled using standard protocols for aldehyde slides. Before and after boiling, slides are incubated with either SyberGreenI, for detection of double stranded DNA, or SyberGreenII, for detection of single stranded DNA (Molecular Probes) and subsequently scanned with a Gene TAC LS IV (Genomics Solutions) to check the transition of double stranded DNA before boiling and single stranded DNA after boiling treatment.

DNA-microarray analysis: the hardware and methods

During the validation of the transcriptome analysis

procedure the following steps have been optimised: RNA isolation, cDNA labeling, and hybridisation.

RNA isolation

RNA isolation is performed by first growing cells in 50 ml GM17 medium and harvesting cells at the desired optical density (OD_{600} ranging from 0.6 to 1.6). From 50 ml MG1363 culture usually 200–300 μg total RNA is routinely obtained. Analysis of isolated RNA is performed using agarose gels in order to assess that the material is of good quality (no degradation, and a 16S to 23S rRNA ratio of approximately 1:2). Alternatively, an Agilent bioanalyzer will be used in the near future, for improved standardisation.

cDNA labeling

Single-strand reverse transcription (amplification) and direct labeling of 25–50 μg of isolated total RNA with Cy3-dCTP or Cy5-dCTP is done either with the Amersham CyScribe First-Strand cDNA Labeling kit or the Invitrogen FluoroScript cDNA labeling system. Both kits give good label incorporation when the labeled cDNA is subsequently used for hybridisation experiments. However, since Cy5 incorporation in cDNA is less efficient than that of Cy3 (due to preference of the reverse transcriptase) and to improve on the sensitivity by enhancing the absolute fluorescence signals, the Amersham CyScribe Post Labelling Kit is currently being tested. Checking the quality of the cDNA is done by hybridising a series of dilutions on a polylysine slide and subsequent analysis with the GeneTAC LS IV.

Hybridisation and scanning

Slides are pre-hybridised in Ambion SlideHyb buffer I for 15 min at 40 °C in a Genomic Solutions Hybstation. After removal of the pre-hybridisation buffer, 5–10 μl of the Cy3/Cy5-labeled cDNA mix in 100 μl Ambion SlideHyb buffer I is added and incubated for 1 h at 42 °C and finally 16 h at 40 °C. Subsequently, the hybridised slides are washed for 1 min in 2× SSC, 0.5% SDS and 5 min in 1×SSC, 0.25% SDS. The slides are scanned using a confocal laser scanner (GeneTAC LS IV).

Signal analysis

After scanning of the slides with the GeneTAC LS IV, spot intensities are determined. The raw data, along with the scanning image are stored in the Molecular Genetics Information System (MolGenIS), which is described below. A grid definition was made to enable the spot analysis software, i.e., Array Pro (Phoretix), to produce tables containing gene names and signal intensities. The program Genespring (Silicon Genetics) is used for further statistical analyses and clustering. The whole procedure for microarray production and analysis is schematically depicted in Figure 1.

DNA-microarray analysis: bioinformatics

Primer design

Large scale production of PCR products of genes or gene fragments is a crucial step in the production of spotted DNA-microarrays. In order to use a high-throughput approach for the production of these PCR products (amplicons) we have developed a software program for the design of primer pairs, named 'GenomePrimer'. A tab-delimited text file or Excel file with the nucleotide sequences of all putative open reading frames (ORFs) is needed as input. GenomePrimer runs under Windows and was written in Borland Delphi 4.0. Various selection parameters for primers, amplicons and annealing temperatures (T_m) can be set within the user-friendly interface.

The parameters to be defined for primer selection are length, GC content, G or C at the 3′ end and inclusion of start or stop codon. For amplicon selection, length, location within an ORF, and flexibility in choice of the length for optimal adjustment of the T_m of the primer can be set. Distribution of G and C and palindromic check of the primer pairs is routinely performed. Calculation of the melting temperature can be performed according to the rules of Suggs et al. (1981) T_m=4(GC)+2(AT) or via the method of Sugimoto et al. (1996), $T_m = 62.3 + 0.41(GC) - 500/length$. Desired tags, for instance for the amplification of all amplicons using one single oligonucleotide pair directed against the 3′ and 5′ tags, can be added to the designed oligonucleotides.

Thousands of primers can be selected within seconds. General result-statistics such as success rate, average primer size, number of nucleotides to synthesize and specific characteristics such as the number of failed primers, (short and/or low GC) amplicons, or number of failed amplicons are given. If primers have not been selected for all ORFs, the setting can easily be adjusted and a new selection run can be performed

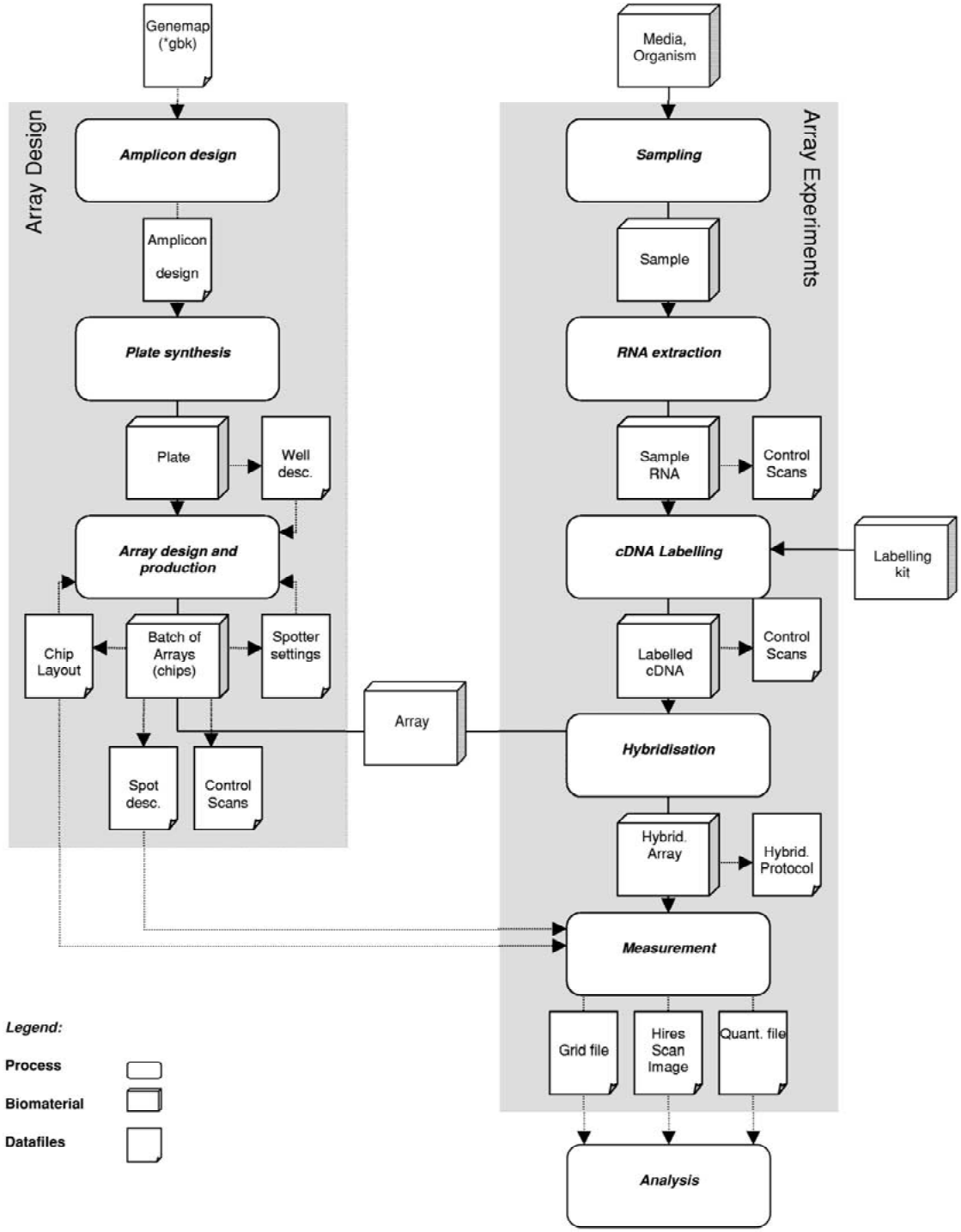

Figure 1. Process overview of a DNA-microarray experiment.

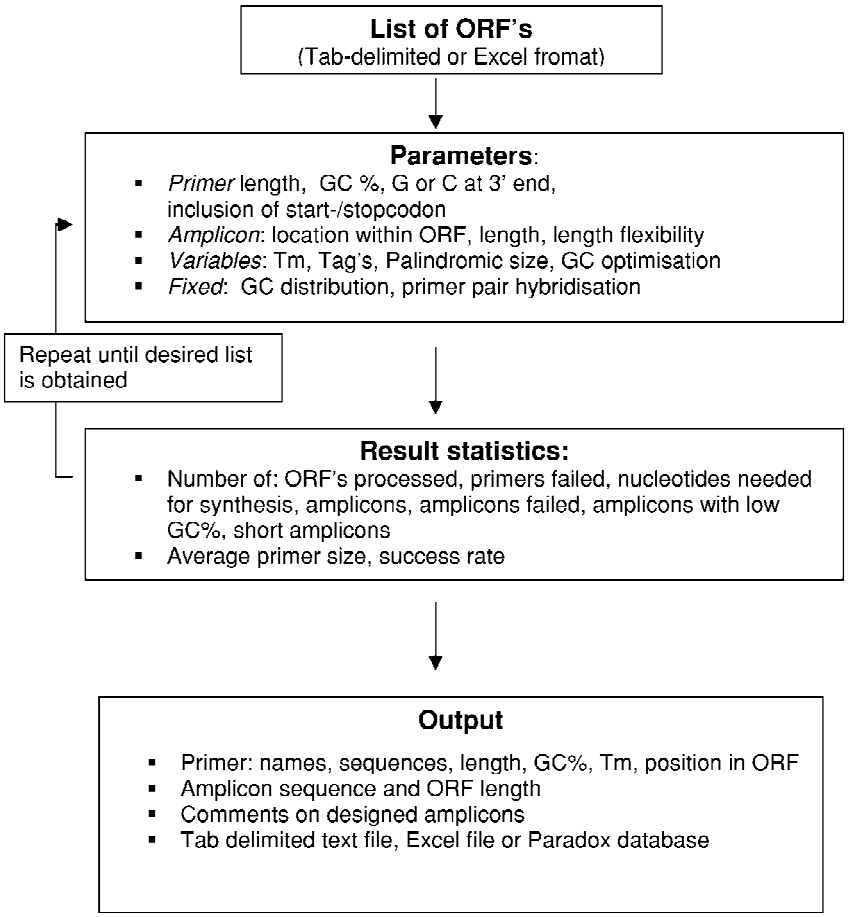

Figure 2. Flow chart of GenomePrimer.

until all primers/amplicons fit within the desired range. The procedure is schematically depicted in Figure 2. The primer pairs and amplicons of the output file can be used as input for the 'Genome2D' program (see below) to indicate their positions within the ORFs.

GenomePrimer was first tested on coding sequences of the lactococcal phage r1t (Table 1). A high-throughput approach, using 96-well microtiter-plates, is used for amplicon production using a pipetting robot in combination with a PCR machine for 96-well plates. After obtaining a 100% success rate for amplicon production with r1t, the program was used to select primers for the complete genome of *S. pneumoniae* TIGR4 and *L. lactis* IL1403, which also gave good results. Simultaneously, a similar software tool for design of primer pairs for DNA microarray construction named 'PrimeArray' was designed by researchers of the Max-Planck Institute of Biology in Switserland (Raddatz et al. 2001). Both GenomePrimer and PrimeArray need the input of coding sequences. Recently the program GST-PRIME was published which can be used to design a large number of primer pairs starting from a list of protein sequences (Varotto et al. 2000). DNA sequences corresponding to the amino acid sequence of the proteins are retrieved, extracted and assembled into gene sequences with and without introns followed by design of primers.

Genome2D: Visualisation of transcript profiles on a linear chromosome map

To visualise a bacterial genome with all its individual genes on a computer screen, we developed a program called Genome2D. Full-screen visualisation of a complete genome enables quick identification of biologically relevant information (Figure 3). For example, it can illustrate which genes have the same orientation and could be transcriptionally linked. Furthermore,

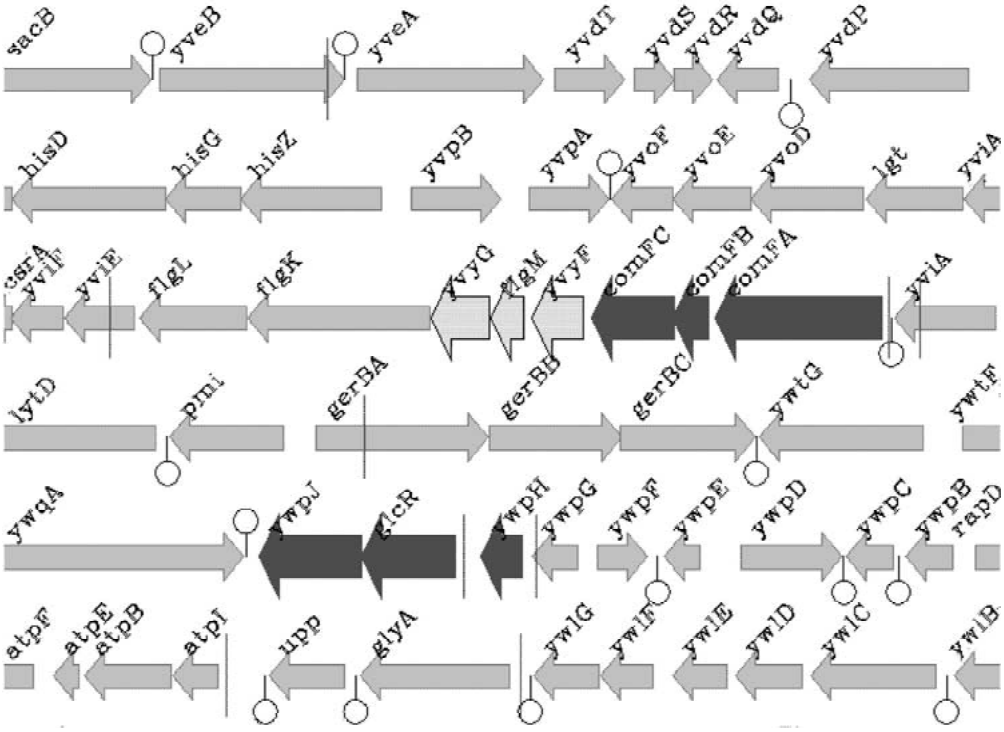

Figure 3. Genome2D visualisation of transcriptional variation between *Bacillus subtilis* 8G5 wild-type and *comK* mutant cells (data from Hamoen et al. manuscript in preparation). The figure displays a partial, but detailed view of the *B. subtilis* genes. Variations in gene expression are indicted by larger arrows (i.e., genes with a 1.5–2× higher expression in WT compared to *comK* cells are coloured *lighter gray*, whereas more than 2× up-regulation is indicated in *darker gray*). Putative terminators shown as stem-loop structures, and predicted *comK* boxes as vertical lines (Hamoen et al. 1998).

Table 1. Overview of primer selection using GenomePrimer and results obtained for amplicon production (GenomePrimer software settings are indicated in grey)

	Phage r1t	*L. lactis* IL1403	S. pneumoniae TIGR4
Nr ORF's	51	2126	2229
GC% ORF's	35.5%	35.5%	40.6%
Primer length (in nt)	17 or 20	18 to 22	18 to 22
Primer GC%	6 to 16	9 to 11	9 to 11
Amplicon length (in bp)	80 to 2000	80 to 800	200 to 800
Amplicon location	Most unique	Most unique	150 bp from ORF ends
Minimal T_m	52 °C	52 °C	56 °C
T_m calculation	T_m=62.3+0.41 (GC)-500/length	T_m=62.3+0.41 (GC)-500/length	T_m=62.3+0.41 (GC)-500/length
Success primer selection	100%	96.5%	98%
Amplicon length (in bp)	89 to 1842	83 to 798	79 to 800
Success rate original PCR	100%	100%	99.5%
Success rate re-amplification	100%	96.2%	100%

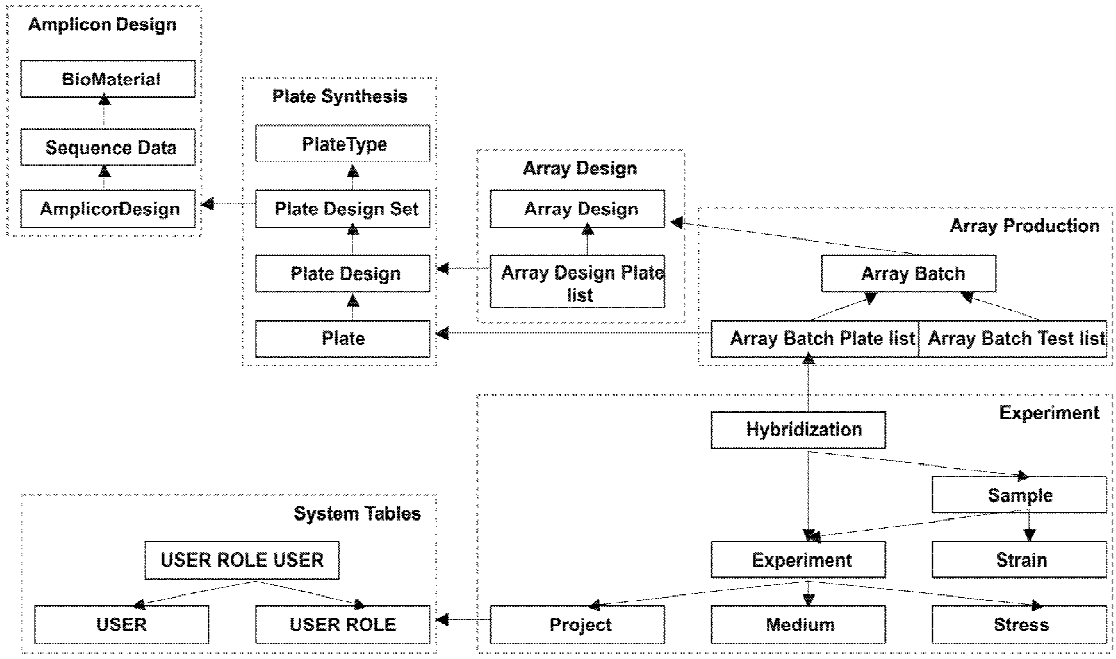

Figure 4. Schematic overview of the MolGenIS information system.

transcriptional terminators or regulator binding sites can be indicated to help in the discovery of operons. Using a simple input file (i.e., a tab-delimited text file, comprising one column with the genes to be colored and one column indicating their colors), subsets of genes can be differentially visualised for DNA-microarray analysis. The input file is based on the differences in transcript levels. This feature enables easy and rapid identification of genes which are transcriptionally linked. In a multiple transcriptome analysis experiment (e.g., measurements during a time course), all data-sets can be loaded as separate input files and subsequently shown in animation. Thus, changes in gene expression levels can be readily recognised.

Genome2D toolbox

Apart from visualisation properties, Genome2D includes a large number of utilities. These comprise tools to: (1) convert gene lists (tab-delimited text files) to FastA format, (2) perform a BLAST analysis at a location of choice (local, NCBI, etc.), (3) search in the genome for binding sites of transcriptional regulators (with or without a trained set), (4) calculate codon usage (of a complete genome or a subset of genes), (5) randomise the genome for statistical purposes (gene fragments or single basepairs), and (6) extract non-coding regions from a genome. Additional utilities can be easily included in the program.

Genome2D development

Genome2D is programmed in Borland Delphi 4. It runs on PCs with Microsoft Windows 98 and higher. To enable the 2D visualisation of a genome, Delphi makes use of the CADSys 4 library version 4.2 (obtained from Piero Valagussa). This library extends the Delphi vectorial graphics support to include 2D/3D CAD-like functions in applications.

MolGenIS: an information system for transcriptome data storage and retrieval

The result of high-throughput DNA-microarray production and analyses will result in large numbers of data. In order to control the processing and storage of these data a dedicated database system, named MolGenIS, has been set-up within the Molecular Genetics group. The MolGenIS system is implemented using the Invengine product of Inventory Netherlands (www.inventory.nl). The setup of an Invengine system is always based on a schematic design file and a user interface design file. MolGenIS is built up of tables that have been grouped in five functional modules: Amplicon design, Plate synthesis, Array design,

Array production and Array experiment (Figure 4). A module for users and user roles has also been added. Access at different restricted levels is given to each user via the use of passwords. The tables of each module have to be completed via a user-friendly interface. Design of the DNA-microarrays and the experiments performed with the arrays are schematically depicted in Figure 1.

Various different types of apparatus/computer-generated files, including data ranging from design to analyses, are stored within MolGenIS. The initial input of the database is a Genbank file map of the genome of bacteria for which DNA-microarrays are being produced. The output file of primers selected by GenomePrimer is stored under 'Amplicon design'. Results of synthesis of the amplicons and their position in the microtiter plates are stored in the 'Well description' file. Under 'Array design and production' the different output files of the software for the GenPack spotter are saved, such as description of the grid-layout of the spotter, spotter settings and the spot description, linking a spot to a microtiterplate well. In addition, a scan of one of the slides from a batch is made after staining with SyberGreen to check the slide quality.

The section 'Array experiments' stores images of the isolated RNA analysed by gel electrophoresis, a scan used to control the labeling of the cDNA by dilution series, and the hybridisation protocol. During measurement of the signal, an ArrayPro grid file with the description of the spot position, which is specific for each type of DNA-microarray, is generated. The results of the scan image obtained with the confocal laser scanner and the ArrayPro software package (image processing solutions), and quantification output files are subsequently generated and stored. The final analysis is performed using the Genespring software package that links the initial input data (Genbank file) with the ArrayPro grid and quantification files. These output results can be linked to pathway data. In addition to the stored files, specific experimental details have to be filled in by the user for each step of the various processes. For instance, within the process of DNA-microarray design and array experiments the following information has to be given: *See table top right*.

Array Design ID	Experiment ID
Spotter Settings (File)	Project ID
Grid Layout (File)	Medium
Spot Notes (File)	Protocol
Log	Temperature
Current date	Shaking
Current user	Stress
	Log
	Current date
	Current user

In a similar way, fields also have to be completed for the description of the project, strain, medium, stress condition, sample size and hybridisation conditions within the DNA-microarray experiment section. In this way all generated data are stored within the MolGenIS database in a highly structured manner, such that it can be easily accessed to compare data obtained from different studies and be made available to other users in any desired format.

LactococCyc: the metabolic pathway database for L. lactis IL1403

The recent availability of the complete chromosomal DNA sequence of *L. lactis* ssp. *lactis* IL1403 enables us to get insight in the numerous properties encoded in this genome (Bolotin et al. 2001; Kuipers 2001). Some 10 years ago, Peter Karp and co-workers started building a knowledge base concerning chemical compounds of intermediary metabolism (Karp 1992). This initiative gradually evolved into the currently well-known encyclopaedia of *Escherichia coli* genes and metabolism (EcoCyc) (Karp et al. 1996). Besides exploration of the metabolic routes of various (at present 11) bacteria in the EcoCyc package (http://biocyc.org/), its computational analysis tool PathoLogic enables users to generate new databases of any sequenced micro-organism of interest. The EcoCyc PathoLogic program uses a genome sequence and annotation in GenBank file format as input. Gene or protein names are extracted/searched (text recognition) from this annotation file and compared to those in the reference database, MetaCyc. Subsequently, when components/enzymes of metabolic routes in the reference database are found, PathoLogic will predict the presence of this metabolic pathway in the analysed organism (Karp et al. 2002).

In a research program on Gram-positive bacteria, we and our collaborators at the Wageningen Center for Food Sciences (Wageningen, The Netherlands) and the Center of Molecular and Biomolecular Informatics (Nijmegen, The Netherlands) aim to reconstruct their metabolic pathways and gene regulatory networks by

performing genome-wide transcriptome analysis of *L. lactis*, *B. subtilis*, *Streptococcus pneumoniae* and *B. cereus* using DNA microarrays. With the aid of the EcoCyc program, in which transcription profiles can be overlayed onto metabolic maps, we expect to be able to identify metabolic pathway-related changes in gene expression (i.e., data filtering).

We have set-up an EcoCyc Pathway/Genome DataBase (PGDB) of *L. lactis* ssp. *lactis* IL1403 by using the EcoCyc PathoLogic program, and called it the LactococCyc PGDB. The program was fed with the GenBank accession-file AE005176. PathoLogic predicted 134 metabolic routes in the genome of IL1403 (whereas 162 pathways are described in EcoCyc for *E. coli*). Out of the 134 predicted metabolic pathways, 88 pathways contained at least 50% of the contributing/participating enzymes in *L. lactis* (for the remaining 46 pathways less than 50% but at least one of the enzymes involved were present). In more detail, from 566 enzymes involved in the various pathways, 408 are present in the genome of IL1403, of which 71% are possibly involved in multiple routes. The MetaCyc reference database (version 5.7) contains 445 pathways composed of 4221 enzymatic and transport reactions. The LactococCyc PGDB comprises 669 enzymatic and transport reactions, while 579 metabolic conversions are missing. This means that either the corresponding enzymes are not present at all in the genome of *L. lactis* IL1403, or the program failed to recognise them, or the genes were not correctly annotated in the GenBank input file. An initial requirement for successful building of a PGDB is that the genome has been adequately annotated, as it is the primary and only input of the PathoLogic program (Ouzounis & Karp 2002). Secondly, since the PathoLogic program produces a prediction of the metabolic content of an organism, the generated PGDB has to be updated with knowledge from literature.

Although the generated *L. lactis* IL1403 PGDB is not fully complete, this does not hamper visualisation of its metabolic pathways and exploration of these with transcriptome data. Reactions missing in pathways are easily recognised, since the corresponding genes (differently colored) are absent in the graphs. Furthermore, the PathoLogic program generates hypertext markup language files displaying the missing reactions or absent genes or gene products. To enable refining and curation of the database an Editor program is included. We are currently curating and refining the PGDB of *L. lactis* IL1403, and building a PGDB of *L. lactis* ssp. *cremoris* strain MG1363.

Concluding remarks

To be able to generate transcriptome data that are both reliable and reproducible, all procedures for production and analysis have to be standardized and validated. Only when this process has been successfully completed, does it make sense to start doing the actual experiments, and to begin to answer biologically relevant questions. Subarrays and complete arrays of IL1403, as well as subarrays containing 96 amplicons of MG1363 are currently being analysed to generate the first reference sets. The full arrays of MG1363 will become available in autumn 2002, providing us with the unique opportunity of doing comparative transcriptome analyses of these two model strains of *L. lactis*. Moreover, methods for genomotyping using DNA–DNA hybridisation are currently being implemented, to be able to analyse the gene content of other lactococcal species and to be able to determine correction factors for the subsequent transcriptome analysis. Moreover, available and future sequence information on lactococcal plasmids, transposons and phages will be used to further enlarge the set of amplicons already spotted on the two master slides.

First experiments will concentrate on the analysis of gene transcription during growth on different media, providing reference sets for the analysis of modified strains or strains grown under different conditions. Moreover, cross hybridisations between IL1403 and MG1363 will be performed to further assess the possibilities and problems associated with using model slides for analysis of non-isogenic strains.

Great challenges now lie ahead to use transcriptome analysis, if possible in conjunction with proteome analysis, for instance (i) to unravel and visualize the complex gene regulatory networks underlying relevant physiological processes, (ii) to study adaptation to changing or completely different environments, (iii) to investigate possible side-effects of genetic engineering and (iv) to optimize the design of metabolic engineering experiments.

Acknowledgements

We wish to express our gratitude to various funding agencies (NWO-ALW, NWO-Bioinformatics program, NWO-STW, BTS-program, IOP-Genomics program, EU-Express Fingerprints FW5 program) for financial support, and our academic, institutional and industrial collaborators for valuable advice, stimulat-

ing discussions and other ways of support. We thank Wiep-Klaas Smits for making available results as an example of Genome2D in Figure 3.

References

Ball KD & Trevors JT (2002) Bacterial genomics: the use of DNA microarrays and bacterial artificial chromosomes. J. Microbiol. Methods 49: 275–284.

Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD & Sorokin A (2001) The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. Genome Res. 11: 731–753.

Diehl F, Grahlmann S, Beier M & Hoheisel JD (2001) Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. Nucleic Acids Res. 29: e38.

Duwat P, Sourice S, Cesselin B, Lamberet G, Vido K, Gaudu P, Le Loir Y, Violet F, Loubiere P & Gruss A (2001) Respiration capacity of the fermenting bacterium *Lactococcus lactis* and its positive effects on growth and survival. J. Bacteriol. 183: 4509–4516.

Kamb A & Ramaswami M (2001) A simple method for statistical analysis of intensity differences in microarray-derived gene expression data. BMC Biotechnol. 1: 1–8.

Karp PD (1992) A knowledge base of the chemical compounds of intermediary metabolism. Comput. Appl. Biosci. 8: 347–357.

Karp PD, Riley M, Paley SM & Pelligrini-Toole A (1996) EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res. 24: 32–39.

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C & Gama-Castro S (2002) The EcoCyc Database.Nucleic Acids Res. 30: 56–58.

Klaenhammer TR et al. (2002) Discovering lactic acid bacteria by genomics. Antonie van Leeuwenhoek, in press.

Kuipers OP (2001) Complete DNA sequence of *Lactococcus lactis* adds flavor to genomics. Genome Res. 11: 673–674.

Kuipers OP, Buist G & Kok J (2000) Current strategies for improving food bacteria. Res. Microbiol. 151: 815–822.

Lucchini S, Thompson A & Hinton JC (2001) Microarrays for microbiologists. Microbiology 147: 1403–1414.

Oliver DJ, Nikolau B & Wurtele ES (2002) Functional genomics: high-throughput mRNA, protein, and metabolite analyses. Metab. Eng. 4: 98–106.

Ouzounis CA & Karp PD (2002) The past, present and future of genome-wide re-annotation. Genome Biology 3: comment2001.

Paton NW, Khan SA, Hayes A, Moussouni F, Brass A, Eilbeck K, Goble CA, Hubbard SJ & Oliver SG (2000) Conceptual modelling of genomic information. Bioinformatics 16: 548–517.

Raddatz G, Dehio M, Meyer TF & Dehio C (2001) PrimeArray: genome-scale primer design for DNA-microarray construction. Bioinformatics 17: 98–99.

Sebaihia M, Thomson N, Holden M & Parkhill J (2001) Microbial genomics. Dynamic duos. Trends Microbiol. 9: 579.

Suggs SV, Wallace RB, Hirose T, Kawashima EH & Itakura K (1981) Use of synthetic oligonucleotides as hybridization probes: isolation of cloned cDNA sequences for human beta 2-microglobulin. Proc. Natl. Acad. Sci. U.S.A. 78: 6613–6617.

Sugimoto N, Nakano S, Yoneyama M & Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Res. 24: 4501–4505.

Tomita M (2001) Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol. 19: 205–210.

Varotto C, Richly E, Salamini F & Leister D (2000) GST-PRIME: a genome-wide primer design software for the generation of gene sequence tags. Nucleic Acids Res. 29: 4373–4377.

Wei Y, Lee JM, Richmond C, Blattner FR, Rafalski JA & LaRossa RA (2001) High-density microarray-mediated gene expression profiling of *Escherichia coli*. J. Bacteriol. 183: 545–56.

Ye RW, Wang T, Bedzyk L & Croker KM (2001) Applications of DNA microarrays in microbial systems. J. Microbiol. Methods 47: 257–272.

Zheng M, Wang X, Templeton LJ, Smulski DR, LaRossa RA & Storz G (2001) DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. J. Bacteriol. 183: 4562–4570.